Assignment_01

Fatima Arshad

1/23/2022

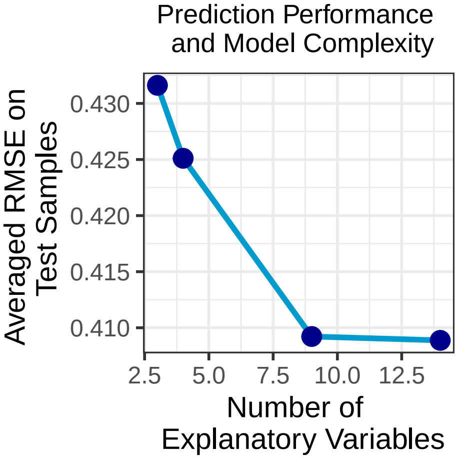
R Markdown

Introduction: This report uses the **computer and mathematical** occupation group from the cps-earnings dataset to build four predictive models using linear regression for earnings per hour. The BIC, RMSE, and five-fold cross-validation for cross-validated RMSE is calculated to discuss the relationship between model complexity and performance.

Exploratory Data Analysis: Data is cleaned and factored before modelling. Age is modeled as its squared non-linearity for a regression with log hourly wage as the dependent variable. The following subset of predictor variables are considered for the purpose of this analysis: education level (Bachelor degree as base), age (squared), sex (male as base), number of children (squared and cubic), and marital status (not married as base). *Refer to Appendix for further details.*

Models: The following four models of varying complexity are defined in this study with Model 1 as the simplest and Model 4 with the highest complexity. 1. **Model 1:** Age and age squared. 2. **Model 2:** Age, age squared, and sex. 3. **Model 3:** Age, age squared, sex, education level, number of children, and number of children squared. 4. **Model 4:** Age, age squared, sex, education level, number of children, number of children squared, married, divorced, interaction term of married and age, interaction term of married and degree higher than bachelors, and interaction term of married and sex.

Model evaluation with the full sample shows that Model 4 is the best according to RMSE and Model 3 is the best according to BIC. Refer to Appendix for further details of each model.



Model 4 performs the best with 5-fold cross-validated RMSE. The graph above suggest that increase in model complexity (by adding more predictor variables) improves performance by lowering the averaged RMSE.However, there is no significant change in model performance by adding more predictor variables and leads to the problem of over-fitting. In conclusion, model selection depends upon finding the best fit while avoiding over-fitting and thus striving for high external validity.

Appendix

Figure 1: Data Filtering and Cleaning The following variables were filtered due to less number of observations:

1. **Marital Status:** 2 (married AF spouse present), 3 (married spouse absent or separated), 4 (widowed), and 6 (separated).
2. **Education (grade92):** Higher education levels are preserved in the data to determine whether it leads to higher wages and following is removed i.e. 32 (1-4th grade), 33 (5th or 6th), 34 (7th or 8th), 35 (9th), 36 (10th), 37 (11th), and 38 (12th grade no diploma).

Figure 2: Skewed Distribution of Hourly Wage

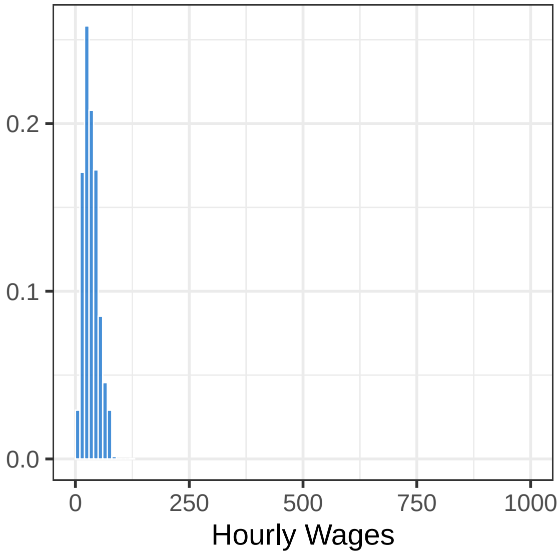


Figure 3: Normal Distribution of Log Hourly Wage

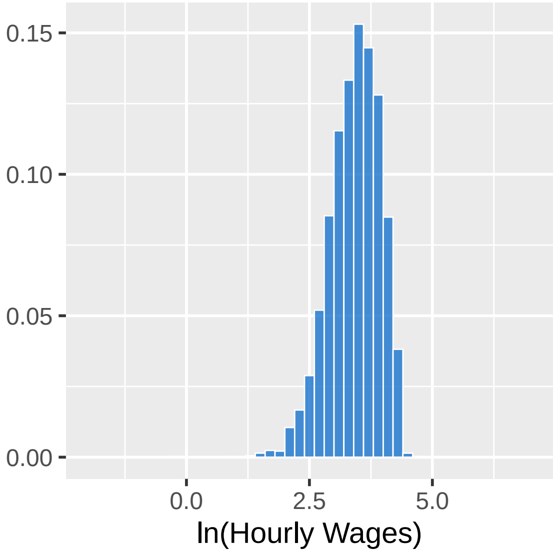
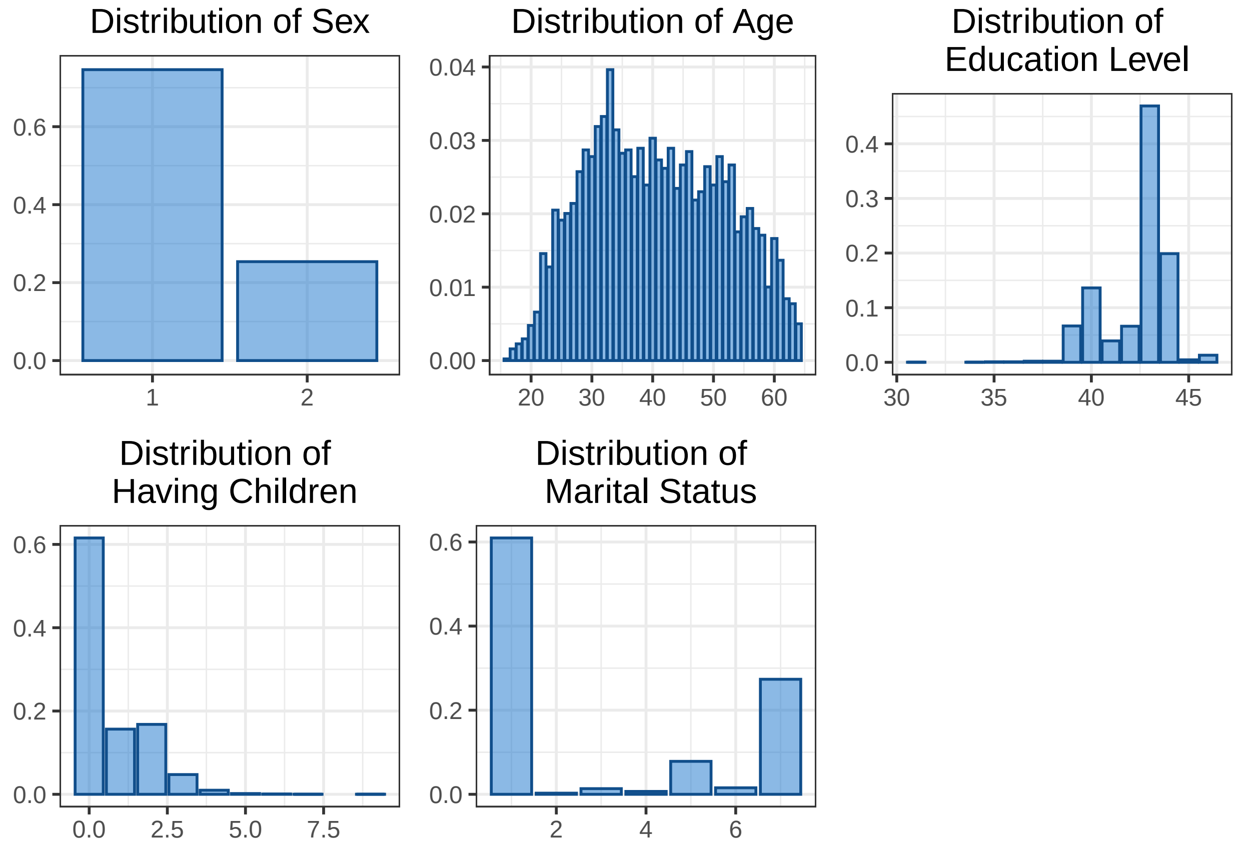


Figure 4: Exploratory Graphs of Predictor Variables



Distribution of Sex

Distribution of Age

Distribution of Education Level

Distribution of Having Children

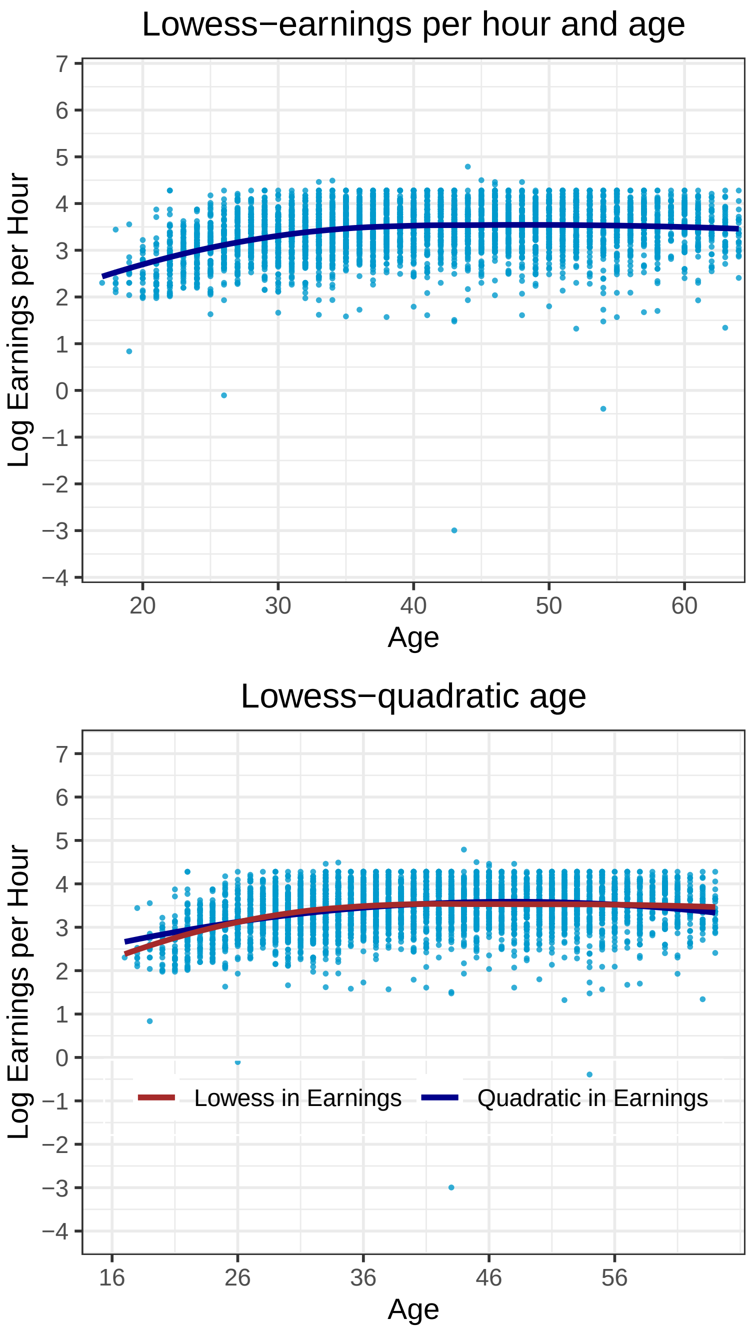
Distribution of Marital Status

Choice of predictor variables:

1. Education Level: Determine if a higher education level in the IT industry leads to more wages.
2. Age: Determine the nature of relationship between age and wages i.e. negative or positive.
3. Sex: Determine the relationship between gender and wages and look out for gender based discrimination.
4. Number of Children: Determine the relationship between number of children and wages because having children increases responsibility on individual.
5. Marital Status: Determine the relationship between marriage and wages especially for females.

Figure 5: Lowess with Observations for Earnings per Hour and Age

```
## `geom_smooth()` using formula 'y ~ x'
```



Lowess−earnings per hour and age

Lowess−quadratic age

Figure 6: Lowess with Observations for Earnings per Hour and Number of Children

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

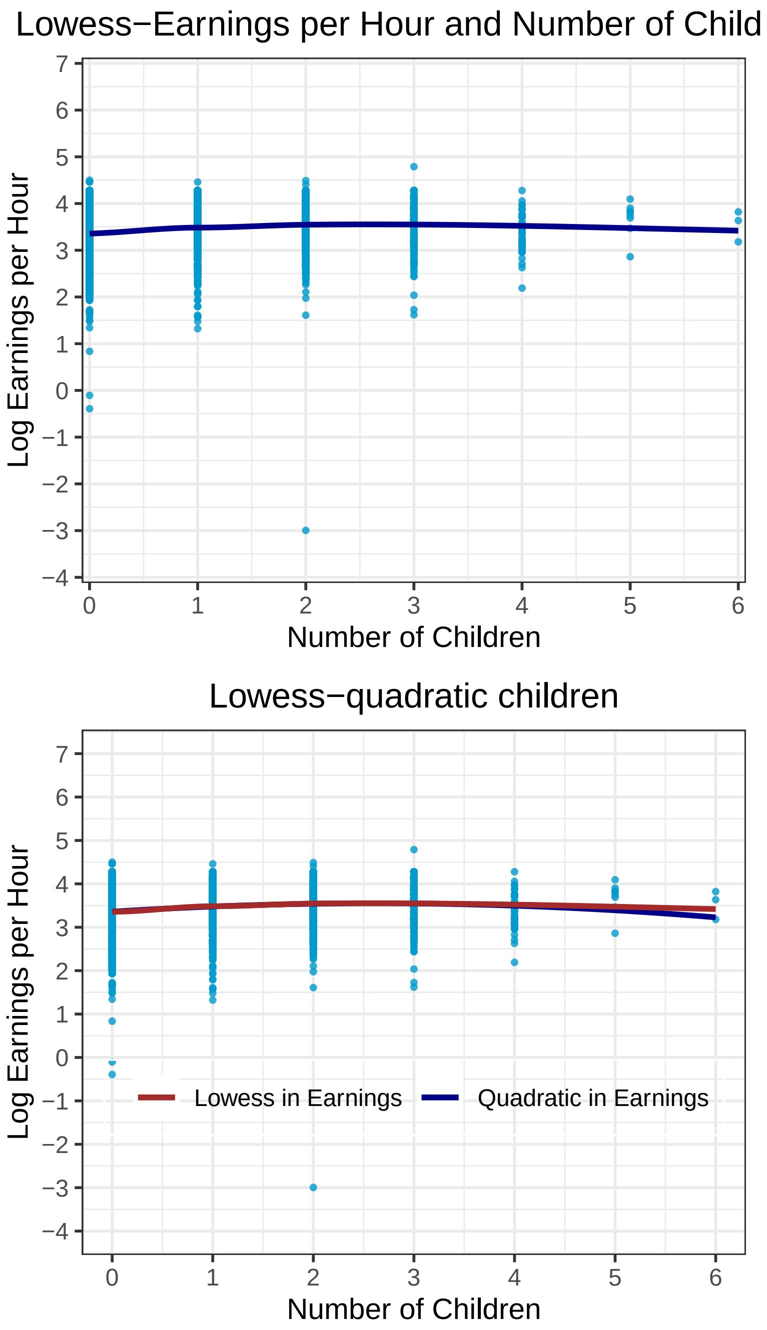


Figure 7: Model Comparison with RMSE and BIC for Full Sample

	(Model1)	(Model2)	(Model3)	(Model4)
Intercept	1.368***	1.395***	1.809***	1.944***
	(0.1023)	(0.1013)	(0.1035)	(0.1349)
Age	0.0927***	0.0929***	0.0721***	0.0681***
	(0.0051)	(0.0051)	(0.0053)	(0.0063)
Age Squared	-0.0010***	-0.0010***	-0.0007***	-0.0007***
	(6.1e-5)	(6.02e-5)	(6.32e-5)	(7.09e-5)
Female		-0.1793***	-0.1769***	-0.1796***
		(0.0180)	(0.0172)	(0.0203)
Higher Than Bachelor			0.1423***	0.1266***
			(0.0176)	(0.0201)
College			-0.2584***	-0.2558***
			(0.0232)	(0.0231)
High School Graduate			-0.2341***	-0.2345***
			(0.0281)	(0.0279)
Number of Children			0.0320*	0.0157
			(0.0151)	(0.0160)
Number of Children Squared			-0.0039	-0.0015
			(0.0041)	(0.0040)
Married				-0.1171
				(0.0721)
Divorced				-0.0761**
				(0.0288)
Age x Married				0.0014
				(0.0019)
Higher Than Bachelor x Married				0.0562
				(0.0397)
Female x Married				0.0244
				(0.0388)
AIC	5,879.9	5,774.8	5,468.8	5,464.5
BIC	5,898.9	5,800.2	5,525.8	5,553.3
RMSE	0.48740	0.48122	0.46343	0.46264
R2	0.12253	0.14466	0.20674	0.20943
Observations	4,194	4,194	4,194	4,194
No. Variables	2	3	8	13

Resample	Model1	Model2	Model3	Model4
Fold1	0.4884910	0.4807721	0.4637807	0.4622138
Fold2	0.4635296	0.4542922	0.4413076	0.4416776
Fold3	0.5015987	0.4919472	0.4735518	0.4727008
Fold4	0.4757802	0.4733453	0.4508029	0.4513919
Fold5	0.5080232	0.5074029	0.4931333	0.4925728
Average	0.4316135	0.4251096	0.4092235	0.4088817

Figure 8: Model Comparison for 5-Fold Cross-Validated RMSE